The secret of my success

Kooi, Barteld; van Ditmarsch, H.P.

Published in:
Synthese

DOI:
10.1007/s11229-006-8493-6

ERRATUM

HANS VAN DITMARSCH and BARTELD KOOI

THE SECRET OF MY SUCCESS

This is a republication of this article, published in *Synthese* (2006) 151: 201–232. Unfortunately many errors occurred due to an electronic flaw. Please find the correct version on the next pages.

Synthese (2006) 153: 339
DOI 10.1007/s11229-006-8493-6

HANS VAN DITMARSCH
and BARTELD KOOI

THE SECRET OF MY SUCCESS

ABSTRACT. In an information state where various agents have both factual knowledge and knowledge about each other, announcements can be made that change the state of information. Such informative announcements can have the curious property that they become false because they are announced. The most typical example of that is 'fact p is true and you don't know that', after which you know that p, which entails the negation of the announcement formula. The announcement of such a formula in a given information state is called an unsuccessful update. A successful formula is a formula that always becomes common knowledge after being announced. Analysis of information systems and 'philosophical puzzles' reveals a growing number of dynamic phenomena that can be described or explained by unsuccessful updates. This increases our understanding of such philosophical problems. We also investigate the syntactic characterization of the successful formulas.

ποταμοῖσι τοῖσιν αὐτοῖσιν ἐμβαίνουσιν ἕτερα
καὶ ἕτερα ὕδατα ἐπιρρεῖ.

Heraclitus
As they step into the same rivers, other and still other waters flow upon them. (D. 12. Translation by C. H. Kahn (Kahn, 1979).)

1. INTRODUCTION

Suppose we discuss New Zealand trees, and I tell you: "You don't know that I have a kowhai tree in my garden". Before I said so, you did not know that I owned such a tree, but after the announcement, that is no longer true: now you *do* know. In the dynamic epistemic logics in Gerbrandy (1999) and van Ditmarsch (2000) this is called an *unsuccessful update*: a formula that becomes false after its announcement. Formally, it is a φ such that $\langle\varphi\rangle\neg\varphi$ is true in some model. Here $\langle\varphi\rangle$ is a dynamic modal operator for the announcement of φ. If atom p describes that I have a kowhai tree in my garden, then $K_{\mathsf{you}}p$ stands for 'You know that p', and the unsuccessful update is $K_{\mathsf{me}}(p \wedge \neg K_{\mathsf{you}}p)$, because $\langle K_{\mathsf{me}}(p \wedge \neg K_{\mathsf{you}}p)\rangle \neg K_{\mathsf{me}}(p \wedge \neg K_{\mathsf{you}}p)$ is true.

Synthese (2006) 151: 201–232
DOI 10.1007/s11229-005-3384-9

There are different logical approaches for reasoning about information change. Besides dynamic epistemic logic, there are most notably belief revision and temporal epistemic logic. The issue of unsuccessful updates does not arise in the belief revision or in temporal epistemic logic, as we will argue here. Therefore we will use dynamic epistemic logic to provide a satisfactory analysis of unsuccessful updates, and use it to analyze problems and puzzles where unsuccessful updates occur.

One of the most influential logical theories about information change is *belief revision* (Alchourrón et al. 1985). It distinguishes three types of information change: expansion, contraction, and revision. Expansion is very much like an announcement, i.e. new information which is consistent with the agent's current information is acquired. A notable difference between AGM expansion and announcements is that success is a postulate for expansion but not a requirement for announcements. Within AGM belief revision this can be achieved because expansion is typically on so-called objective formulas only, i.e., formulas without modal operators. In our current setting, such AGM belief revision corresponds to announcements of facts and their boolean combinations, and in this setting these are not very interesting.

In the theory of belief revision, expansion is fully characterized by six *rationality postulates* (Gärdenfors 1988), and turns out to be set-theoretic union. Consequently, repeated expansion with the same formula has the same effect as expanding once, and the order in which the expansions are executed does not matter. But in the current context these become immediately and unmistakably crucial: for example, after I say "you do not know that I have a kowhai tree in my garden," I cannot say that again: the revision cannot be repeated. And even though I can first say "you can imagine that I do not have a kowhai tree in my garden" and only then say "(but actually I have one in my garden and) you do not know that (I have a kowhai tree in my garden)," I cannot reverse the order of these two announcements: after the last, the first can no longer be made.

There are also temporal epistemic logics for reasoning about information change (Fagin et al. 1995). In these approaches information change occurs as time progresses, however the propositional content of information change cannot be expressed in the logical language. This is because the temporal operators are not in themselves descriptive of the change. For example, your knowledge after my announcement of "you do not know that I have a kowhai tree

in my garden" can be described by $XK_{\mathsf{you}}p$, where X is the 'next'-operator that assumes an underlying transition between an information state before and after the announcement. But the X here does not reveal anything about the nature of the announcement. In other words, the issue of unsuccessful update does not arise in this approach either.

The issue of unsuccessful updates is closely related to Moore's problem (Hintikka 1962; Sorensen 1988), which concerns sentences that can be true, but that cannot be known to someone. These Moore-sentences may be unsuccessful updates. For example, $p \wedge \neg Kp$ is satisfiable, but $K(p \wedge \neg Kp)$ is inconsistent in epistemic logic. The relation with our running example will be obvious. We do not present Moore-problems in this paper separately.

The structure of the paper is as follows. In Sections 2 and 3 we define the logic of public announcements and various notions of successful and unsuccessful formulas and updates. This fine-tuning in terminology and corresponding semantics provides the available tools for the analysis of various problems and puzzles.

Sections 4–7 are entirely devoted to the analysis of problems, or problem areas, that can be resolved by using the notion of unsuccessful updates. Section 4 deals with the Muddy Children problem. The epistemic paradox known as the Surprise Examination paradox is discussed in Section 5. Card games are the subject of Section 6. In Section 7 security protocols are analyzed. In Section 8 we present different attempts to characterize the successful formulas and a small technical contribution to that ongoing discussion.

2. THE LOGIC OF PUBLIC ANNOUNCEMENTS

In this section, we present the logic of public announcements with common knowledge. This logic contains both epistemic and dynamic modal operators. With epistemic operators we express individual knowledge, for an arbitrary agent, and public (common) knowledge, for the entire group of agents. With dynamic modal operators we express the effect of public announcements, i.e., public (and truthful) announcements of formulas in the language. The parameters that play a static role throughout the semantic and syntactic definitions are a set of agents N and a set of propositional atoms P.

An *epistemic model* $M = \langle S, \sim, V \rangle$ consists of a *domain* S of *factual states* or just states, *accessibility* $\sim: N \to \mathcal{P}(S \times S)$ which for each agent $n \in N$ defines a binary accessibility relation (that is an equivalence relation) $\sim_n$ on S, and a *valuation* $V : P \to \mathcal{P}(S)$ which for each atom $p \in P$ defines a valuation $V_p \subseteq S$. The class of epistemic models is named $\mathcal{S}5_N(P)$. If M is an epistemic model, and $s \in \mathcal{D}(M)$ (s is in the domain of M), then the pointed model (M, s) is an *epistemic state*.

The language of public announcements $\mathcal{L}^u_N(P)$ is inductively defined as

$$\varphi \quad ::= \quad p \mid \neg\varphi \mid (\varphi \wedge \psi) \mid K_n\varphi \mid C\varphi \mid [\varphi]\psi$$

We assume the reader to be familiar with the interpretation of the propositional and epistemic operators. Intuitively, $K_n\varphi$ stands for 'agent n knows φ,', and $C\varphi$ stands for 'it is common knowledge that φ.' The construct $[\varphi]\psi$ stands for 'after truthful public announcement of φ, it holds that ψ'. Instead of 'φ is a public and truthful announcement' we say 'φ is an announcement'. The dual of $[\varphi]$ is $\langle\varphi\rangle$, so that $\langle\varphi\rangle\psi$ stands for 'after *some* announcement of φ, it holds that ψ'. As announcements are functional, $\langle\varphi\rangle\psi$ entails $[\varphi]\psi$. Notational abbreviations are defined as usual, including – assuming a finite number of agents – general knowledge $E\varphi$.

The language $\mathcal{L}^u_N(P)$ is interpreted on epistemic models and epistemic states. Given an epistemic model $M = \langle S, \sim, V \rangle \in \mathcal{S}5_N(P)$ and a state $s \in \mathcal{D}(M)$, we define inductively:

$$\begin{array}{ll}
M, s \models p & :\Leftrightarrow s \in V_p \\
M, s \models \neg\varphi & :\Leftrightarrow M, s \not\models \varphi \\
M, s \models \varphi \wedge \psi & :\Leftrightarrow M, s \models \varphi \text{ and } M, s \models \psi \\
M, s \models K_n\varphi & :\Leftrightarrow \text{for all } t \in S : s \sim_n t \text{ implies } M, t \models \varphi \\
M, s \models C\varphi & :\Leftrightarrow \text{for all } t \in S : s \sim_N t \text{ implies } M, t \models \varphi \\
M, s \models [\varphi]\psi & :\Leftrightarrow M, s \models \varphi \text{ implies } M|\varphi, s \models \psi
\end{array}$$

In the clause for $C\varphi$, access $\sim_N$ is defined as the reflexive transitive closure of the union of access for all individual agents, i.e., $\sim_N := (\bigcup_{n \in N} \sim_n)^*$. In the clause for $[\varphi]\psi$, $M|\varphi := \langle S', \sim', V' \rangle$ is defined as follows:

$$\begin{array}{l}
S' := \{s' \in S \mid M, s' \models \varphi\} \\
\sim'_n := \sim_n \cap (S' \times S') \\
V'_p := V_p \cap S'
\end{array}$$

In other words: the model $M|\varphi$ is the model M restricted to all the states where φ holds, including access between states. The interpretation of the dual $\langle\varphi\rangle$ of $[\varphi]$ is obvious: $M, s \models \langle\varphi\rangle\psi$ if and only if $M, s \models \varphi$ and $M|\varphi, s \models \psi$. Validity, and validity in models, are defined as usual. Example interpretations are given in Section 4 on the Muddy Children Problem.

A proof system for this logic is found in Table I. It is a special case of the general proof system for the logic of epistemic actions presented in Baltag et al. (2003), with precursors in Plaza (1989) and Gerbrandy (1999). Soundness and completeness is shown for the general proof system.[1] The principles relating announcements to knowledge are the axiom *announcement and knowledge*, which also (partly) expresses that an announcement is a *partial* function, and the derivation rule *announcement and common knowledge*, which is a recipe to derive common knowledge after an announcement. Other valid principles include:

$[\varphi](\psi \to \chi) \to ([\varphi]\psi \to [\varphi]\chi)$	normality for $[\varphi]$
$C(\varphi \to E\varphi) \to \varphi \to C\varphi$	induction
$\langle\varphi\rangle\psi \to [\varphi]\psi$	functionality of announcement
$[\varphi]\psi \leftrightarrow (\varphi \to [\varphi]\psi)$	truthful announcement

TABLE I

The logic of public announcements

All propositional tautologies	
$\vdash K_n(\varphi \to \psi) \to K_n\varphi \to K_n\psi$	normality of K_n
$\vdash K_n\varphi \to \varphi$	truth axiom
$\vdash K_n\varphi \to K_nK_n\varphi$	positive introspection
$\vdash \neg K_n\varphi \to K_n\neg K_n\varphi$	negative introspection
$\vdash C(\varphi \to \psi) \to C\varphi \to C\psi$	normality of C
$\vdash C\varphi \to (\varphi \wedge EC\varphi)$	use of C
$\vdash [\varphi]p \leftrightarrow (\varphi \to p)$	announcement and atoms
$\vdash [\varphi]\neg\psi \leftrightarrow (\varphi \to \neg[\varphi]\psi)$	announcement and negation
$\vdash [\varphi](\psi \wedge \chi) \leftrightarrow ([\varphi]\psi \wedge [\varphi]\chi)$	announcement and conjunction
$\vdash [\varphi]K_n\psi \leftrightarrow (\varphi \to K_n[\varphi]\psi)$	announcement and knowledge
from $\vdash \varphi$ and $\vdash \varphi \to \psi$ follows $\vdash \psi$	modus ponens
from $\vdash \varphi$ follows $\vdash K_n\varphi$	necessitation for K_n
from $\vdash \varphi$ follows $\vdash C\varphi$	necessitation for C
from $\vdash \varphi$ follows $\vdash [\psi]\varphi$	necessitation for $[\psi]$
from $\vdash \chi \to [\varphi]\psi$ and $\vdash \chi \wedge \varphi \to E\chi$ follows $\vdash \chi \to [\varphi]C\psi$	announcement and common knowledge

Some of these can be derived schematically, while others need an induction on one of the occurring formula variables. For example, both 'induction' is a schematic consequence of the 'rule of announcement and common knowledge'. The logic of public announcements is not a normal modal logic, because uniform substitution is not a derivation rule: the whole point of this investigation is, that even though $[q]q$ is valid, $[p \wedge \neg Kp](p \wedge \neg Kp)$ is *in*valid! The logic of public announcements is decidable, see Baltag et al. (2003). For details on the overview in this section, see van Ditmarsch et al. (2005).

3. SUCCESSFUL FORMULAS

In this section, we define successful and unsuccessful formulas by their semantic properties. The logic of public announcement with common knowledge provides a sufficient logical context for these definitions.

Let us recapitulate once more the communicative expectations, and how we can be so thoroughly deceived by these. If an agent truthfully announces φ to a group of agents, it appears *on first sight* to be the case that (s)he 'makes φ common knowledge' that way: in other words, if φ holds, then after announcing that, $C\varphi$ holds. In other words, that $\varphi \rightarrow [\varphi]C\varphi$ is valid. We have seen in the introduction that this expectation is unwarranted, because the truth of epistemic parts of the formula may be influenced by its announcement. But sometimes the expectation *is* warranted after all: formulas that always become common knowledge after being announced, will be called *successful*. We can also distinguish various degrees of success, depending on the context of an epistemic state. Let us see what the possibilities are.

After announcing φ, φ sometimes remains true and sometimes becomes false, and this depends both on the formula *and* on the epistemic state. The introductory example involved an epistemic state for one atom p and two agents, from now on called Anne and Bill, where Anne knows the truth about p but Bill doesn't. This epistemic state is formally defined as $(Letter, 1)$, where model $Letter$ has domain $\{0, 1\}$, accessibility relation for agent a is $\sim_a := \{(0, 0), (1, 1)\}$ (that is: the identity on the domain), accessibility relation for agent b is $\sim_b := \{(0, 0), (1, 1), (0, 1), (1, 0)\}$ (that is: the universal relation on the domain), and valuation $V_p = \{1\}$. The model

is called *Letter* because it can be seen as the result of (only) Anne reading a *letter* which contains the truth about p.

If in this epistemic state $(Letter, 1)$ Anne says, truthfully: "I know that p," then after this announcement $K_a p$, it *remains true* that $K_a p$:

$$Letter, 1 \models [K_a p] K_a p$$

This is, because in *Letter* the formula $K_a p$ is true in state 1 only, so that the model $Letter|K_a p$ consists of the singleton state 1, with reflexive access for a and b. It also becomes common knowledge that Anne knows p: we have that $Letter, 1 \models [K_a p] C K_a p$; although in this particular case of the singleton model $(Letter|K_a p)$, a description involving common knowledge is not very informative.

But it is not always the case that announced formulas remain true. In the given epistemic state $(Letter, 1)$, Anne could on the other hand have said as well, to Bill: "You don't know that p." The actual implicature in this case is "Fact p is true and you don't know that." After this announcement $K_a(p \wedge \neg K_b p)$, that also only succeeds in state 1, Bill knows that p, therefore $K_a(p \wedge \neg K_b p)$ is *no longer* true

$$Letter, 1 \models [K_a(p \wedge \neg K_b p)] \neg K_a(p \wedge \neg K_b p)$$

and so it is certainly not commonly known.

The epistemic state transition induced by this update is visualized in Figure 1. In the visualization, we link states that are the same for an agent and label the link with the agent's name, and we assume reflexivity and transitivity of access. Please remember these conventions; the following pictures will be more complex, and, unlike here, transitivity of access may then play a part in the visualization.

Note that in this particular epistemic state, announcement of $K_a p$ induces the same state transition as announcement of $K_a(p \wedge \neg K_b p)$. The first remains true, but the second becomes false. For a given state transition we can always find a formula that induces it and remains true. We will address that matter in Section 8.

Incidentally, $[K_a(p \wedge \neg K_b p)] \neg K_a(p \wedge \neg K_b p)$ is even valid: the announced formula will always become false, *if* the announcement can

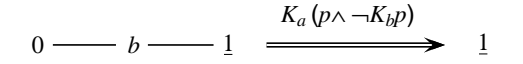

Figure 1. A simple unsuccessful update: Anne says to Bill "(p is true and) you don't know that p."

be executed at all. Mostly for the intuitive and motivating appeal of formulas that 'become common knowledge', i.e. the successful ones, we will focus on that, rather than on the 'always unsuccessful' ones, that will therefore not be defined or considered separately.

In between these extremes of 'always successful' and 'always unsuccessful' there are also formulas that sometimes remain true, and at other times – given other epistemic states – become false after an announcement. An example of that is the (implicit) announcement of 'nobody knows whether (s)he is muddy' in the Muddy Children problem, to be discussed in detail in Section 4. This formula is successful for all rounds except the round where the muddy children learn that they are muddy. Does this mean that 'successful' is not a fixed property of a formula, but that it is relative to an epistemic state? High time for further precision.

DEFINITION 1 (*Successful formula*). A formula $\varphi \in \mathcal{L}_N(P)$ is successful if and only if $[\varphi]\varphi$ is valid. A formula is *unsuccessful* if and only if it is not successful.

DEFINITION 2 (*Successful update*). Given a formula $\varphi \in \mathcal{L}_N(P)$ and an epistemic state (M, s).

- φ is successful in (M, s) if and only if $M, s \models \langle\varphi\rangle\varphi$
- φ is unsuccessful in (M, s) if and only if $M, s \models \langle\varphi\rangle\neg\varphi$.

In the first case φ is a *successful update* (in that epistemic state), and in the last case that φ is an *unsuccessful update* (in that epistemic state).

In the definitions, the switch between the 'box' and the 'diamond' versions of the announcement operator may puzzle the reader. In the definition of a successful *formula* we really need the 'box'-form: clearly $\langle\varphi\rangle\varphi$ is invalid for all φ except tautologies. But in the definition of a successful *update* we really need the 'diamond'-form: clearly, whenever the announcement formula is false in an epistemic state, $[\varphi]\neg\varphi$ would therefore be true. That would not capture the intuitive meaning of an unsuccessful update, because that is formally represented as a feature of an epistemic state transition. We must therefore assume that the announcement formula can indeed be truthfully announced. This explains the difference between the two definitions.

Updates with true successful formulas are always successful, but sometimes updates with unsuccessful formulas are successful. By 'always' ('sometimes') we mean 'in all (some) epistemic states'. The truth of the *first* will be obvious: if a successful formula φ is true in an epistemic state (M, s), then $\langle\varphi\rangle\varphi$ is also true in that state, so it is also a successful update. A typical example of the *last*, to be explained in detail later, is the already mentioned action of 'not stepping forward' in Muddy Children, that was said to be 'only unsuccessful in the last round and otherwise successful'. We can actually distinguish different degrees of 'success', that will also nicely match somewhat tentative distinctions made in the literature:

DEFINITION 3 (*Individually/generally/publicly successful*). Given a formula $\varphi \in \mathcal{L}_N(P)$ and an epistemic state (M, s).

- φ is *individually successful* in (M, s), or *successful for agent n* in (M, s), if and only if $M, s \models \langle\varphi\rangle K_n\varphi$
- φ is *generally successful* in (M, s) if and only if $M, s \models \langle\varphi\rangle E\varphi$.
- φ is *publicly successful* in (M, s) if and only if $M, s \models \langle\varphi\rangle C\varphi$.

Similarly, we define individually, generally, and publicly *unsuccessful*, e.g., φ is individually unsuccessful if $M, s \models \langle\varphi\rangle K_n\neg\varphi$, etc.

Individually unsuccessful corresponds to $\langle\varphi\rangle K_n\neg\varphi$ and not to $\langle\varphi\rangle\neg K_n\varphi$. The *first* expresses that agent n knows φ to be false after it has been announced. This is stronger than φ 'merely' being false after the announcement: that may or may not be known to the agent. Whereas the *last* already follows if φ is an unsuccessful update: if φ is false after its announcement, this obviously cannot be known to any agent. Nothing is therefore to be gained by such a distinction. The other group notions of unsuccessful are similarly motivated. Note that publicly successful implies generally successful implies individually successful implies successful (update), but not the other way round. Similarly, publicly unsuccessful implies generally unsuccessful implies individually unsuccessful implies unsuccessful (update), but not the other way round. In particular, successful updates may well be publicly unsuccessful. We did not make a similar distinction for validities for a simple reason:

PROPOSITION 4. Let $\varphi \in \mathcal{L}_N(P)$. Then $[\varphi]\varphi$ is valid if and only if $[\varphi]C\varphi$ is valid.

Proof. Completeness allows for a very short proof: '$\vdash [\varphi]\varphi$ implies $\vdash [\varphi]C\varphi$' is an instance of the rule for announcement and common knowledge, for $\chi := \top$ and $\psi := \varphi$. The direction '$\vdash [\varphi]C\varphi$ implies $\vdash [\varphi]\varphi$' rather trivially follows from $\vdash C\varphi \rightarrow \varphi$ (use of C), followed by $[\varphi]$-necessitation and $[\varphi]$-normality, plus some propositional reasoning and MP.

COROLLARY 5. Let $\varphi \in \mathcal{L}_N(P)$. All the following validities are equivalent: $[\varphi]\varphi$, $[\varphi]K_n\varphi$ for some agent n, $[\varphi]E\varphi$, $[\varphi]C\varphi$.

So, for validities, the four notions of successful ('as such', individually, generally, publicly) all coincide, but for formulas in general, they do not. In particular, $[\varphi]\varphi$ is *not* logically equivalent to $[\varphi]C\varphi$. The distinction is also useful, because it appears not to be made in some relevant literature, in particular not in the original publication (Gerbrandy 1999, pp. 100–101), Gerbrandy takes 'individually successful' as the primitive notion for both successful updates and successful formulas.

The following makes precise that the successful formulas 'do what we want them to do': if true, they become common knowledge when announced.

COROLLARY 6. $[\varphi]\varphi$ is valid if and only if $\varphi \rightarrow [\varphi]C\varphi$ is valid.

Which formulas are successful? The syntactic characterization of successful formulas will be addressed in Section 8. An answer to this question is not obvious, because some inductive ways to construct the class of successful formulas fail: even if φ and ψ are successful, $\neg\varphi$, $\varphi \wedge \psi$, or $\varphi \rightarrow \psi$ may be unsuccessful.

Before we present our partial results towards characterization of the successful updates, we show in detail concrete examples of unsuccessful updates in relevant communicative settings. This forms the main part of our contribution. It illustrates the relevance of our subject matter to philosophical analysis and the analysis of multi-agent systems.

4. MUDDY CHILDREN

The Muddy Children puzzle is one of the best known puzzles that involve knowledge. It is known that versions of this puzzle were

circulating in the fifties. The earliest source of the puzzle we could find is a puzzle book by Gamow and Stern (1958). They present the 'cheating wives' version.

> The great Sultan Ibn-al-Kuz was very much worried about the large number of unfaithful wives among the population of his capital city. There were forty women who were openly deceiving their husbands, but, as often happens, although all these cases were a matter of common knowledge, the husbands in question were ignorant of their wives' behavior. In order to punish the wretched women, the sultan issued a proclamation which permitted the husbands of unfaithful wives to kill them, provided, however, that they were quite sure of the infidelity. The proclamation did not mention either the number or the names of the wives known to be unfaithful; it merely stated that such cases were known in the city and suggested that the husbands do something about it. However, to the great surprise of the entire legislative body and the city police, no wife killings were reported on the day of the proclamation, or on the days that followed. In fact, an entire month passed without any result, and it seemed the deceived husbands just did not care to save their honor.
>
> "O Great Sultan," said the vizier to Ibn-al-Kuz, "shouldn't we announce the names of the forty unfaithful wives, if the husbands are too lazy to pursue the cases themselves?"
>
> "No," said the sultan. "Let us wait. My people may be lazy, but they are certainly very intelligent and wise. I am sure action will be taken very soon."
>
> And, indeed, on the fortieth day after the proclamation, action suddenly broke out. That single night forty women were killed, and a quick check revealed that they were the forty who were known to have been deceiving their husbands. (Gamow and Stern 1958, pp. 20–21).[2]

A version with cheating husbands rather than wives can be found in Moses et al. (1986). Another version features wise men (McCarthy 1990). The version that is most popular today, and the version we discuss in this paper, involves muddy children. It was introduced in Barwise (1981). Given a group of children whose forehead may be muddy or not, and who can only see the foreheads of other children, a father announces that at least one of them is muddy, and that those children who are sure whether their own forehead is muddy, should step forward. If there are n children, then the n-th time he announces this, all the muddy children step forward.

The first analysis of this problem with epistemic logic was presented in Moses et al. (1986), where the dynamics were modelled on the metalevel rather than in the logical language. The first object level treatment of this problem using the logic of public announcements was by Plaza. (He called this logic the logic of public communications and used slightly different notation.) An analysis of

the problem using dynamic epistemic logic can be found in Jelle Gerbrandy's dissertation (Gerbrandy 1999).

The Muddy Children puzzle is a classic example of unsuccessful updates: from the announcement that nobody knows whether he or she is muddy, the muddy children may learn that they are muddy. Although not stepping forward is strictly speaking not a public announcement, because the children do not make an utterance, pragmatically it is an announcement of their ignorance.

To give the general idea of the analysis using public announcement logic, we look at the special case of three children Anne, Bill, and Cath (a, b, and c). Let us suppose two of the children, Anne and Bill, are muddy. After two announcements the muddy children know that they are muddy. Intuitively this can be seen as follows. Suppose you are Anne. Then you see that Bill is muddy and Cath is not. Now the father says at least one of the children is muddy. From this you can infer that if you are not muddy, Bill cannot see anyone who is muddy, and therefore he would infer he is muddy. So if the father now asks those children who know whether they are muddy to step forward, and none step forward, you infer that Bill did not know he was muddy. The only explanation for this is that you are muddy yourself. Therefore, the next time the father asks you to step forward if you know whether you are muddy, you step forward. The situation is the same for Bill. Therefore he steps forward too.

We can represent the initial situation of the (three) muddy children problem with a cube. Each of the three children can be muddy or not. For this we introduce three propositional variables: m_a, m_b, m_c. Therefore there are eight possible states. In the general case for n children, we get an 'n-hypercube', where the planes correspond to the muddiness of the children. We call the model for three children $Cube$. A picture of this model is shown in the top of Figure 2. The states are labelled xyz, where $x, y, z \in \{0, 1\}$, where $x = 1$ means that Anne is muddy, and $y = 0$ means that Bill is not muddy, etc. In state 110, for instance, a and b are muddy, but c is not. Let us assume that 110 is the actual state. Although it is the case in 110 that everybody knows there is at least one muddy child this is *not* common knowledge. For example a considers it possible that b considers it possible that no child is muddy. Formally, we have that $Cube, 110 \models E(m_1 \vee m_2 \vee m_3)$, but that $Cube, 110 \models \neg C(m_1 \vee m_2 \vee m_3)$, because $110 \sim_a 010 \sim_b 000$ and $Cube, 000 \models \neg(m_1 \vee m_2 \vee m_3)$.

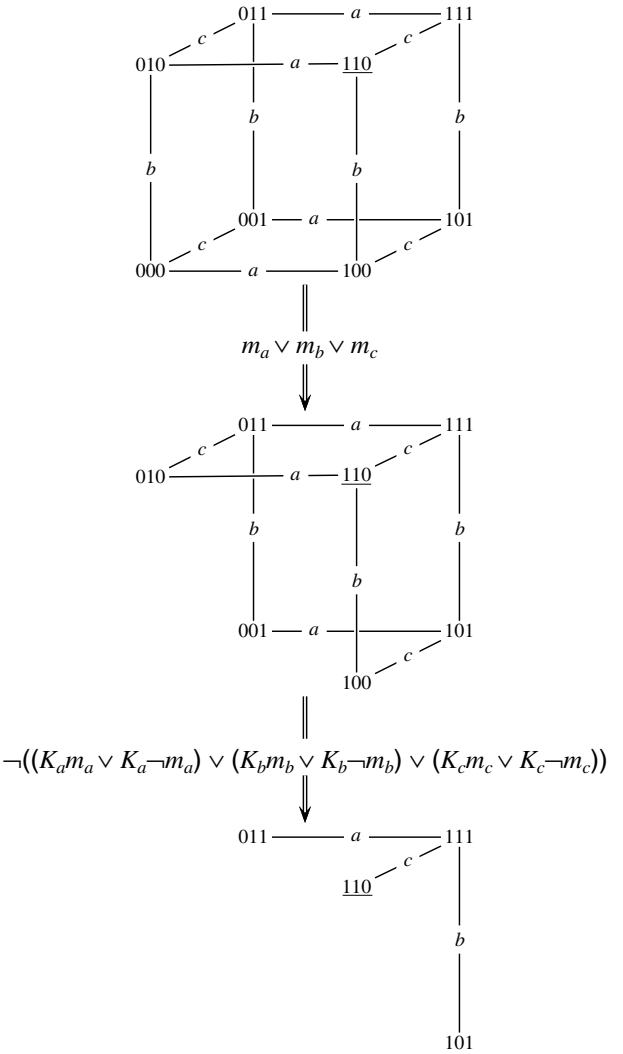

Figure 2. The top depicts the epistemic state where Anne and Bill are muddy, and Cath is not. The first transition depicts the effect of the announcement that at least one child is muddy. The second transition depicts the effect of no child stepping forward, i.e., an announcement that nobody knows whether (s)he is muddy. This is an unsuccessful update. At the bottom epistemic state, Anne and Bill know that they are muddy.

In Figure 2 we have visualized the changes of information for that situation. The first announcement of the father says that at least one of the children is muddy: $m_a \vee m_b \vee m_c$. Let us abbreviate this formula with muddy:

$$\mathsf{muddy} = m_1 \vee m_2 \vee m_3$$

The formula muddy is false in the state 000, therefore by the semantics of public announcements presented in Section 2, we get a new model where this state is eliminated, and the accessibility relations are adapted accordingly. This is the model in the middle of Figure 2. One can simply check that $Cube|\mathsf{muddy}, 110 \models C\mathsf{muddy}$. Therefore muddy is successful in $(Cube, 110)$. But it is not only successful in *Cube*, it is a successful *formula*, i.e. $[\mathsf{muddy}]\mathsf{muddy}$ is valid. After the announcement that at least one child is muddy, at least one child is muddy.

The epistemic state we have thus acquired, has a special feature. When one focuses on the state where exactly one child is muddy, one sees that each of these states is indistinguishable from another state for only two of the children. That means that one child knows what the actual state would be if there were only one muddy child. In particular, the child who knows what the actual state would be, knows he is muddy. For instance:

$$Cube|\mathsf{muddy}, 100 \models K_a m_a$$

Now the father asks those children who know whether they are muddy to step forward. When no one steps forward, this means that no child knows whether he or she is muddy. The formula that expresses that at least one child knows whether he or she is muddy, is $\mathsf{knowmuddy}$ (unlike 'knowing that φ', $K\varphi$, 'knowing whether φ' is described by $K\varphi \vee K\neg\varphi$).

$$\mathsf{knowmuddy} = (K_a m_a \vee K_a \neg m_a) \vee (K_b m_b \vee K_b \neg m_b) \vee (K_c m_c \vee K_c \neg m_c)$$

The announcement made by the children not stepping forward is $\neg\mathsf{knowmuddy}$. Consequently, those states where exactly one child is muddy are eliminated by this announcement. By using the semantics of Section 2 we get the bottom epistemic state shown in Figure 2.

This last formula is unsuccessful in $Cube|\mathsf{muddy}, 110$. The announcement that none of the children know whether they are

muddy or not yields a situation where two children do know that they are muddy. Before the announcement, both children know that if he is not muddy, then the other child knows he is muddy. By learning that none of the children know, they can infer that they must be muddy themselves. More formally:

$$Cube|\mathsf{muddy}, 110 \models \langle \neg \mathsf{knowmuddy} \rangle \mathsf{knowmuddy}$$

So $\neg$knowmuddy is unsuccessful in $(Cube|\mathsf{muddy}, 110)$. Is is also individually unsuccessful for Anne and Bill in that epistemic state, but not for Cath. So it is therefore not generally unsuccessful. Formally:

$$Cube|\mathsf{muddy}, 110 \models \langle \neg \mathsf{knowmuddy} \rangle K_a \mathsf{knowmuddy}$$
$$Cube|\mathsf{muddy}, 110 \models \langle \neg \mathsf{knowmuddy} \rangle K_b \mathsf{knowmuddy}$$
$$Cube|\mathsf{muddy}, 110 \models \langle \neg \mathsf{knowmuddy} \rangle \neg E \mathsf{knowmuddy}$$

However, when all three children are muddy, $\neg$knowmuddy is a successful update:

$$Cube|\mathsf{muddy}, 111 \models \langle \neg \mathsf{knowmuddy} \rangle \neg \mathsf{knowmuddy}$$

This is because in that state of the model, even at the bottom of Figure 2, every child still considers it possible he or she is not muddy.

The Muddy Children puzzle teaches some important lessons. Firstly, it can be informative to announce something that everybody knows. Every child knows at least one child is muddy, but the announcement does give the children information. The fact becomes common knowledge. The same formula can have different amounts of success depending on the context. Moreover, as the example of three muddy children showed, repetitions of the same announcement can be informative.

5. SURPRISE EXAMINATION

The Surprise Examination paradox has a relatively short history of about 60 years. Apparently the Swedish mathematician Lennart Ekbom heard a message on the radio during World War II announcing a civil defense exercise, which was to take place in the next week. It was also announced that this exercise would be a surprise.

Then he noticed that there was something paradoxical about this announcement. Kvanvig (1998) and Sorensen (1988, pp. 253). The paradox was first published by O'Conner (1948):

> Consider the following case. The military commander of a certain camp announces on a Saturday evening that during the following week there will be a "Class A blackout". The date and time of the exercise are not prescribed because a "Class A blackout" is defined in the announcement as an exercise which the participants cannot know is going to take place prior to 6.0 p.m. on the evening in which it occurs. It is easy to see that it follows that the exercise cannot take place at all. It cannot take place on Saturday because if it has not occurred on the first six days of the week it must occur on the last. And the fact that the participants can know this violates the condition which defines it. Similarly, because it cannot take place on Saturday, it cannot take place on Friday either, because when Saturday is eliminated Friday is the last available day and is, therefore, invalidated for the same reason as Saturday. And by similar arguments, Thursday, Wednesday, etc., back to Sunday are eliminated in turn, so that the exercise cannot take place at all. (O'Connor 1948)

One of the first replies to this was that the exercise could take place after all:

> Suppose that the Commanding Officer arranges for a blackout to take place during the period covered by the announcement. Clearly the date of its occurrence cannot be forecast from the announcement. So it will by definition be a Class-A blackout, and he will be entirely justified in his announcement that a Class-A blackout would take place during this period. (Scriven 1951)

There are many versions of this paradox. There is one involving a prisoner that is sentenced to death by hanging on an unexpected day, which is why the paradox is also known as the Hangman Paradox. This version was introduced in Quine (1953). The version which is most popular nowadays, introduced in Weiss (1952), involves a surprise exam. A teacher announces to his students there will be an exam next week, but the exact day of the exam will be a surprise. We will study this version here. The first analysis of the Surprise Examination paradox with dynamic epistemic logic was done by Gerbrandy in his dissertation. We follow his analysis.

The first step of the analysis is to formalize the utterances of the teacher. Let us take as the set of propositional variables the set {mo, tu, we, th, fr}. The propositional variable "mo", for instance, means that the exam will take place on Monday. The announcement of the teacher that there will be an exam next week, can easily be formalized as

$$\begin{aligned}\mathsf{exam} = {} & (\mathrm{mo} \wedge \neg \mathrm{tu} \wedge \neg \mathrm{we} \wedge \neg \mathrm{th} \wedge \neg \mathrm{fr}) \vee \\ & (\neg \mathrm{mo} \wedge \mathrm{tu} \wedge \neg \mathrm{we} \wedge \neg \mathrm{th} \wedge \neg \mathrm{fr}) \vee \\ & (\neg \mathrm{mo} \wedge \neg \mathrm{tu} \wedge \mathrm{we} \wedge \neg \mathrm{th} \wedge \neg \mathrm{fr}) \vee \\ & (\neg \mathrm{mo} \wedge \neg \mathrm{tu} \wedge \neg \mathrm{we} \wedge \mathrm{th} \wedge \neg \mathrm{fr}) \vee \\ & (\neg \mathrm{mo} \wedge \neg \mathrm{tu} \wedge \neg \mathrm{we} \wedge \neg \mathrm{th} \wedge \mathrm{fr})\end{aligned}$$

This is an exclusive disjunction over the possible days.

Gerbrandy distinguishes two readings of the second sentence that is announced by the teacher:

1. *Given the information the students now have*, the students will not know the day of the exam in advance.
2. The students will not know the day of the exam in advance, *even after they hear this announcement*.

The first sentence can be formalized in the language of the logic of public updates. The second one however cannot be formalized using public announcement logic, because of the self-reference that is involved in it. Let us start with a formalization of the first reading. Here we take Anne as a representative of the students:

$$\begin{aligned}\mathsf{surprise} = {} & \mathrm{mo} \rightarrow \neg K_a \mathrm{mo} \wedge \\ & \mathrm{tu} \rightarrow [\neg \mathrm{mo}] \neg K_a \mathrm{tu} \wedge \\ & \mathrm{we} \rightarrow [\neg \mathrm{mo}][\neg \mathrm{tu}] \neg K_a \mathrm{we} \wedge \\ & \mathrm{th} \rightarrow [\neg \mathrm{mo}][\neg \mathrm{tu}][\neg \mathrm{we}] \neg K_a \mathrm{th} \wedge \\ & \mathrm{fr} \rightarrow [\neg \mathrm{mo}][\neg \mathrm{tu}][\neg \mathrm{we}][\neg \mathrm{th}] \neg K_a \mathrm{fr}\end{aligned}$$

The idea is a follows. First, if the exam is on Monday, she does not know it. Next, after school has finished on Monday, if the exam is on Tuesday, then late on Monday (when she has learned the exam is not on Monday) she does not know the exam will be on Tuesday, and so on.

The question is what happens to Anne's information state when the announcements are made. One can view these as two separate announcements: first exam and then $\mathsf{surprise}$. If one insists that all updates are successful, then one is committed to saying that $[\mathsf{exam}][\mathsf{surprise}] K_a(\mathsf{exam} \wedge \mathsf{surprise})$. The first thing to note is that $K_a(\mathsf{exam} \wedge \mathsf{surprise})$ is inconsistent, because of the *reductio ad absurdum* that eliminates all the days step by step. Suppose it is consistent. The following reasoning can be done from Anne's point of view. Suppose exam is true. Therefore $(\mathrm{mo} \vee \mathrm{tu} \vee \mathrm{we} \vee \mathrm{th} \vee \mathrm{fr})$. We proceed by cases. Suppose that the exam is on Friday, i.e. fr is true. From exam it follows that $(\neg \mathrm{mo} \wedge \neg \mathrm{tu} \wedge \neg \mathrm{we} \wedge \neg \mathrm{th} \wedge \mathrm{fr})$. Then it

follows from surprise that $[\neg mo][\neg tu][\neg we][\neg th]\neg K_a fr$. But we assumed that $(\neg mo \wedge \neg tu \wedge \neg we \wedge \neg th \wedge fr)$. Therefore these announcements can be executed. The result of their execution is an epistemic state where $K_a fr$ is true, because only states where fr is true remain. This contradicts the previous. So fr cannot be the case. Anne also arrives at this conclusion therefore $K_a \neg fr$. Using this we can now proceed in the same way to derive $K_a \neg th$ and so on. Together this contradicts $K_a$exam. This is not paradoxical, it simply means that surprise is not successful for Anne.

Let us now look in detail at what happens to an epistemic state that represents the situation when Anne is completely ignorant about the truth values of the propositional variables, and first learns exam. The resulting model has five states. In each of these states exactly one propositional variable is true. Let us assume that the exam is actually on Monday. This is the state on the left in Figure 3. It is easy to see that in this state (when the exam is on Monday) the last four conjuncts of surprise are trivially true, because the antecedents are false. The first conjunct is true, because Anne does not know the exam will be on Monday.

When we look at the other states in this model we see that the only state where surprise is false, is the state where the exam is on Friday. In that world fr is true and

$$[\neg mo][\neg tu][\neg we][\neg th]K_a fr$$

also holds, because Friday would be the only remaining day Anne considers to be possible. Therefore

$$fr \rightarrow [\neg mo][\neg tu][\neg we][\neg th]\neg K_a fr$$

is false and so surprise is false in this state. So the state where fr is true is eliminated. That shows, that when surprise is announced, Anne learns the exam will take place on another day than Friday. The result of the announcement surprise yields the second model of Figure 3.

So the only thing that the students learn from the second announcement is that the exam will not take place on Friday, but the *reductio ad absurdum* cannot go any further, because the announcement is not individually successful. Anne does not know that the day of the exam will be a surprise after the announcement. If, however, the teacher repeats his announcement, Anne would be able to eliminate another day, and this continues as long as the

teacher can (truthfully) repeat it, as is shown in Figure 3. Assuming the exam takes place on Monday, the teacher cannot repeat his announcement after the fourth time: ⟨surprise⟩¬surprise is true in the fourth model of Figure 3. In the first three epistemic states the announcement is successful (after the announcement the exam would still be a surprise, given Anne's information at that point), although in all first four epistemic states the announcement is not individually successful.

The problem that still remains is another reading of the teacher's second announcement "The students will not know the day of the exam, *even after they hear this announcement*." It is clear that this announcement is self-referential. Many analyses of the liar paradox blame the paradox on its self-referential nature. This also seems to be the case here, the addition of "even after they hear this announcement" forces the update to be successful, i.e. it forces the children to conclude after the announcement that they will not know the day of the exam in advance. This is inconsistent, as was noted earlier, and it seems that as in the case of the liar, we can blame the self-reference of the announcement.

6. CARD GAMES

Just as the Surprise Examination paradox, the Card Games example is rooted in a World War II civil defense exercise. The game of

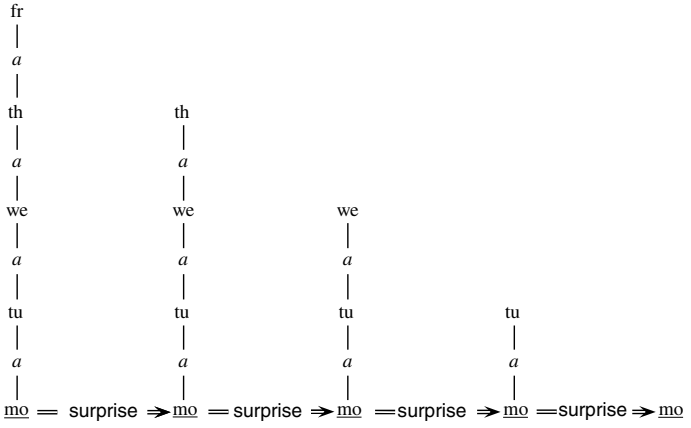

Figure 3. Updates for the Surprise Examination paradox when the teacher's announcement is repeated and the exam is on Monday.

Cluedo was invented by Elva and Anthony E. Pratt. Anthony Pratt supposedly came up with the idea for the game while on nightly fire-patrol during blackouts, in the Birmingham area in 1943. This 'murder-mystery' game, first published in the late 1940s, has been hugely popular for over 50 years. 'Normally' Cluedo players gain information, and win, because they exchange factual information (confirm or deny ownership of cards), but it turns out that a fully *rational* Cluedo player, i.e. a perfectly logical reasoner only acting in his own interests, may win because other rational players *implicitly* announce epistemic information, namely that they cannot win (van Ditmarsch 2002). One may think of such an action both as an involuntary action resulting in a loss, such as a 'forced move resulting in mate' in chess, or as a voluntary action actively harming one's interests, such as 'scoring an own goal' in soccer. First, we explain a bit more about the game.

A murder is committed. The player who finds out who the murderer is, what the murder weapon was, and in which room the murder was committed, wins the game. The game is played on a game board with a picture of the murder house, with nine rooms in it and 'paths' leading in a certain number of steps from one room to another. There are six players. There are six guest cards, six weapon cards, and nine room cards. The three categories of cards are shuffled separately. One suspect card, one weapon card and one room card are blindly drawn and put apart. These 'murder cards' represent the actual murderer, the murder weapon and the murder room. All remaining cards are shuffled together. They are then dealt to the players. Every player gets three cards.

A player's move consists of the following: *Throw the dice*. Try to *reach a room* by walking your pawn over the game board. The number of steps on the board may not exceed the outcome of the throw of dice. If a room is reached *voice a suspicion* about it, i.e. about a guest, a weapon and that particular room. As a consequence of the suspicion, the pawn with the same colour as that of the suspected player is moved to the suspected room, and that weapon token is placed in that room. *Gather responses* to that suspicion from the other players. The other players respond to the suspicion in clockwise fashion: either a player doesn't have any of the requested cards, he says so, and the next player responds to the suspicion; or a player holds at least one of the requested cards, he shows exactly one of those to the requesting player *only*, and no further responses may be gathered. You may now either *end your move* (who is next in turn

is again determined clockwise) or, if you think you know enough, *make an accusation*. An accusation is also the combination of a suspect, a weapon and a room card, but it plays another role in the game than a suspicion does: Each player can make an accusation only once. It is not voiced but written down. The accusing player then checks the three murder cards, without showing them to others. If the accusation is false, that player has lost and the game continues. If the cards match the accusation, it is successful. The first player who makes a successful accusation, wins the game.

The logical description of some of these game moves requires a dynamic epistemic logic that is slightly more involved than the logic of public announcements: the action of showing a card to another player with the remaining players 'looking on' is more complex than a public announcement (Gerbrandy 1999; van Ditmarsch 2000; Baltag et al. 2003). We merely highlight the implicit move that takes place when you pass your turn to the next player: for a perfectly rational player this means that you are announcing that you cannot win the game *yet*, or in other words, that you do not know what the murder cards are. It is conceivable that this is so informative to other players, that they therefore win if they are now to move, even before asking a single question. And more than that: that in a situation where nobody can win, because somebody announces that, somebody (else) can win: an unsuccessful update! An example where the opponents gain factual knowledge from a given player' inability to win, is the following: suppose that in the first move of the game, Anne voices the suspicion 'Green has done it with a knife in the ballroom'. Nobody shows a card. If (and only if) Anne now does *not* make a final accusation, the other players can conclude that those three cards cannot be the murder cards and that Anne must hold at least one of them. In other words, they learn facts about cards. Learning the murder cards, and then winning, is just learning facts about very specific cards.

As the models and actions for the full game of Cluedo are rather complex, we present the unsuccessful update by means of a simpler example, namely only three players each holding one card. Assume that the players are Anne, Bill, and Cath (a, b, c), and that they each hold one of the cards $0, 1, 2$. The epistemic model describing this situation consists of six card deals. For example, 012 describes the deal where (in that order) Anne holds card 0, Bill holds card 1, and Cath holds card 2; card deals where a player holds the same card, are indistinguishable for that player. For example, Anne

cannot distinguish 012 from 021. This induces obvious equivalence relations for all players on this model. We call the model *Hexa* (see Figure 4).

Assume that in epistemic state $(Hexa, 012)$ an outsider tells the three players that the deal of cards is neither 201 nor 120. This corresponds to the announcement $\neg(2_a \wedge 0_b \wedge 1_c) \wedge \neg(1_a \wedge 2_b \wedge 0_c)$. Abbreviate $Hexa|(\neg(2_a \wedge 0_b \wedge 1_c) \wedge \neg(1_a \wedge 2_b \wedge 0_c))$ as $Hexa'$. In $Hexa'$ none of the agents know that the deal is 012, but all three consider it possible for the other two to know it (see Figure 4). If Anne now announces that she doesn't know Bill's card, then Cath can derive Bill's card from that announcement. Formally

$$Hexa', 012 \models [\neg(K_a 0_b \vee K_a 1_b \vee K_a 2_b)](K_c 0_b \vee K_c 1_b \vee K_c 2_b)$$

The corresponding unsuccessful update actually is, that after Anne says that she considers it possible that no player knows the card deal, she no longer considers that possible, because she now knows that both Bill and Cath know the card deal.

In full precision: Let δ_d be an atomic description of a card deal (characteristic function of the state valuation), e.g. $\delta_{012} := 0_a \wedge \neg 1_a \wedge \neg 2_a \wedge \neg 0_b \wedge 1_b \wedge \neg 2_b \wedge \neg 0_c \wedge \neg 1_c \wedge 2_c$. Then $\mathsf{win}_n := \bigvee_{d \in D} K_n \delta_d$ describes that agent n knows the deal of cards, where D is the current set of states (card deals) in the domain, and $\mathsf{win} := \bigvee_{n \in N} \mathsf{win}_n$ describes that some agent can win. The unsuccessful update just described is formally

$$Hexa', 012 \models \langle \neg K_a \mathsf{win} \rangle K_a \mathsf{win}$$

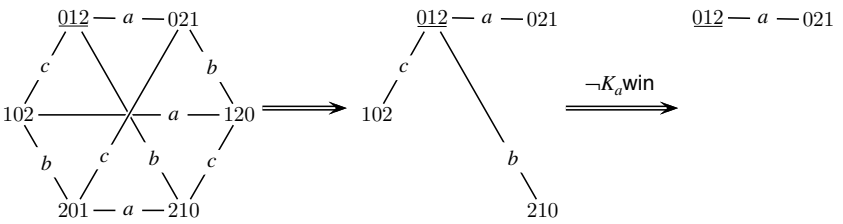

Figure 4. The effect of two announcements on the epistemic state for a card deal where Anne holds 0, Bill holds 1, and Cath holds 2. The first transition, unlabelled, depicts the effect of $\neg(2_a \wedge 0_b \wedge 1_c) \wedge \neg(1_a \wedge 2_b \wedge 0_c)$ ('the deal is neither 201 nor 120'). After Anne tells the others that she cannot win, both Bill and Cath can win. In the form of an unsuccessful update: After Anne announces that she does not know if some player can win, she knows that some player can win.

although the conceptually more appealing description (that is not an unsuccessful update) of Cath winning because Anne announces that she cannot, is formally

$$Hexa', 012 \models \langle \neg \mathsf{win}_a \rangle \mathsf{win}_c$$

Both yield the same transition for this epistemic state. We have visualized the unsuccessful update in Figure 4. Another unsuccessful update in $Hexa'$, inducing a different transition, is $\neg\mathsf{win}$: this formula only succeeds in point 012; therefore in the resulting singleton epistemic state everybody can win!

In Cluedo, the goal is not full knowledge of the card deal, but only partial knowledge, namely of the ownership of the murder cards. The implicit action of announcing that you cannot win appears not to have been noted before (van Ditmarsch 2002). It is highly relevant from a game theoretical perspective. Cluedo is a game of imperfect information, where players' optimal strategies depend on their opponents' strategies: Nash equilibria determine what is optimal. Because 'can't win' actions must also be modelled as part of players' strategies, they will also determine what is optimal. Specifically, strategies that were thought to have been optimal, may well turn out to be suboptimal when 'can't win' actions are also taken into account. In plain words: suppose you're in the heat of a Cluedo game, have just arrived in the ballroom, and have to decide on the best question to ask. It may well be that without 'can't win' actions the best question to ask is – incorrectly – 'Green did it with a knife in the ballroom,' but that *with* these actions the best question to ask is 'Scarlett did it with a rope in the ballroom'. Whether such situations can really occur, is not yet known to us.

7. SECURITY PROTOCOLS

This application is about communicating agents ('sender' and 'receiver') that try to keep the content of their communications from eavesdroppers that are listening in. The details are of a rather combinatorial character, which will therefore not be presented in great detail. The rough idea is, that the information that is to be conveyed is 'weakened' by presenting this in the presence of various alternatives. As long as the receiving party has enough informational advantage over the eavesdropper(s), the communication can be successful while the secret is kept. The topic is partly rooted in 19th

century born design theory (Kirkman 1847), a subdiscipline of combinatorial mathematics that investigates how many different ways there are to convey information, such as these secrets. Its interest to the philosophical community is mainly due to the curious way in which pragmatics and semantics are mixed up: publicly known intentions of agents involved in protocols become part of the meaning of their statements, in other words: the pragmatics are drawn into the semantics. We regard this as highly relevant for the analysis of knowledge and belief change. By way of the semantic modelling of such intentions, we can explain scenarios in which they are 'self-defeating' in the precise sense of an unsuccessful update.

The specific setting is the 'seven cards' or 'Russian cards' problem. This was first posed in the 2000 Moscow Mathematics Olympiad, to which it was suggested by A. Shapovalov. An extensive analysis of its epistemic aspects can be found in van Ditmarsch (2003), see also Makarychev and Makarychev (2001).

Given are three players, Anne, Bill, and Cath, and a pack of seven known cards, 0, 1, 2, 3, 4, 5, 6. Anne and Bill each draw three cards and Cath gets the remaining card. Anne and Bill want to openly (publicly) inform each other about their cards, without Cath learning from any of their cards who holds it.
Assume that Anne has drawn 0, 1, and 2, and Bill 3, 4, and 5, so that Cath gets card 6. Anne now says, publicly: "I have $\{0, 1, 2\}$, or I don't have any of these cards," after which Bill says, also publicly: "Cath has card 6."

These two announcements do not solve the problem. Using Anne's intention to keep her cards a secret from Cath, Anne actually reveals all her cards to Cath. Why is that so, and why does it *appear* to solve the problem?

Obviously, Bill immediately derives Anne's hand from her announcement: if Anne had none of 0, 1, and 2, Bill should have had at least two of those. But he does not. Also, Cath seems unable to pin down any specific card on Anne from that announcement, e.g., both 012 and 345 still appear possible hands for Anne: if she held 345, indeed that includes none of 0, 1, and 2. Beyond that, the underlying protocol seems 'safe' enough from Anne's point of view: if she had said instead 'I do not have card 6' that would not have been informative to Cath either, but rather risky for Anne, who does not know at this stage whether Cath holds card 6 or not, and *if* Cath had held card 5, she would have derived from Anne's announcement that Bill holds 6: a loss again. Further, Bill's announcement informs

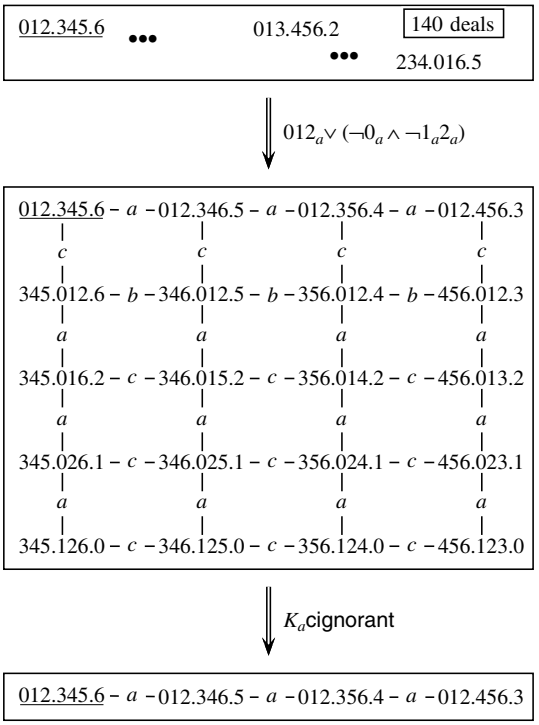

Figure 5. The results of Anne's announcement and of her intention. The second transition depicts the unsuccessful update: note that $K_a$cignorant is false in the final state.

Anne of Bill's cards, namely the remaining three, and Cath already knew which card she held, so that doesn't help her much either.

To understand what is wrong completely, we resort to a formal analysis. The underlying models are completely analogous to those for Cluedo, and to *Hexa*:

We call the model where the cards have been dealt: *Rus*. A deal of cards where Anne holds 0, 1, and 2, Bill holds 3, 4, and 5, and Cath holds 6, is represented by 012.345.6, etc. The epistemic model representing the information players have about each other, is defined using the same modelling principles as in *Hexa*: it is commonly known what the deck of cards is, how many cards each 'player' holds, and that players only know their own hand of cards. Anne's first announcement is described by the formula

$$012_a \vee (\neg 0_a \wedge \neg 1_a \wedge \neg 2_a)$$

Abbreviate this as announce. Cath's ignorance about Anne's and Bill's cards is described by the formula

$$\bigwedge_{q=0\ldots6} \neg K_c q_a \wedge \neg K_c q_b$$

After an update of $(Rus, 012.345.6)$ with announce we reach the information state pictured in Figure 5. One can easily check in the figure that Anne indeed *knows* that Cath is ignorant of her cards after her announcement. (All of the four deals in the first row, that represent 'what Anne knows given the actual deal', have a single alternative for Cath, namely the deal in the row directly beneath it, in which the hands of Anne and Bill are swapped: so Cath cannot determine ownership of any of those cards.) However, Cath doesn't know *that*, and, surprisingly, Cath can derive factual knowledge from that ignorance. Let us first walk through the figure, to establish that Cath does not know that Anne knows that Cath is ignorant: or, in other words, to establish that Cath can imagine that Anne can imagine that Cath knows.

For Cath, after the update with announce, deal 012.345.6 can still not be distinguished from deal 345.012.6 (below 012.345.6 in the figure). But in 345.012.6 Anne does not know that Cath is ignorant: in that case, Anne would have considered it possible that the deal was, e.g., 345.016.2 (below 345.012.6 in the figure). And if the deal had been 345.016.2, Cath would have known that Bill has cards 0 and 1 (as this holds for all four deals in the row containing 345.016.2, in the figure).

Why is this informative for Cath? Cath rightfully assumes that Anne wouldn't *dare* make an announcement that might inform her. In other words, it is not merely required that Cath is ignorant after Anne's announcement, but also that Anne knows that Cath is ignorant, 'and this is commonly known to all'. Another way of saying that, is that Anne 'really' says: "announce is true, and after having said that, Cath still doesn't know my cards." Formally: announce $\wedge$ [announce]$K_a$cignorant. This is equivalent to the sequence of two announcements announce and $K_a$cignorant (see the derived principles in Section 2). Restriction of the model resulting from announce to the states where $K_a$cignorant is true, retains only the top four states in the figure (see the transition there). And in *that* state Cath knows the entire card deal. So cignorant is false, and a fortiori

$K_a$cignorant is false too. In an overview (for convenience we stick to the box-version of the dynamic operators):

$Rus, 012.345.6 \models$ [announce $\wedge$ [announce]$K_a$cignorant]$\neg K_a$cignorant
$Rus, 012.345.6 \models$ [announce][$K_a$cignorant]$\neg K_a$cignorant
$Rus|$announce$, 012.345.6 \models [K_a$cignorant]$\neg K_a$cignorant

The second update, on $K_a$cignorant, is merely explicitizing Anne's intentions. Without those intentions, just updating with announce would have sufficed. This would have happened if an outsider – having no interest in keeping or communicating secrets – had said "Anne has either hand $\{0, 1, 2\}$, or none of those cards." So, Cath only learns Anne's cards from Anne's intention to prevent Cath learning her cards. Without that intention, Cath would not have learnt Anne's cards.

By now the reader may wonder how the Russian Cards Problem can be solved anyway. For that, think, or see the references! There are various non-trivial solutions.

8. SYNTACTIC CHARACTERIZATION OF SUCCESSFUL UPDATES

We have defined the successful formulas, and the various related notions, *semantically*. What is the fragment of the logical language that consists of the successful formulas? We have not fully answered this question yet, nor, as far as we know, has this been answered by others. Public announcement logic is decidable, so one can also decide whether a formula is successful or not, but we would like to be able to say so in a straightforward way only using the syntactic form of the formula. In this section, we report on our progress and the results of other researchers.

Let us start with some negative results. Even when φ and ψ are successful, their conjunction may be unsuccessful: For example, $p \wedge \neg K_n p$ is unsuccessful, but both p and $\neg K_n p$ are successful. That may be not immediately clear for the last case, therefore we present the – short – proof:

Let M, s be arbitrary. We have to prove that $M, s \models [\neg K_n p]\neg K_n p$, in other words, that $M, s \models \neg K_n p$ implies $M|\neg K_n p, s \models \neg K_n p$. Let $M, s \models \neg K_n p$. Then there must be a $t \sim_n s$ such that $M, t \models \neg p$, and therefore also $M, t \models \neg K_n p$, and therefore $t \in M|\neg K_n p$. From $s \sim_n t$ in $M|\neg K_n p$ and $M|\neg K_n p, t \models \neg p$ follows $M|\neg K_n p, s \models \neg K_n p$.

Also φ may be successful but $\neg\varphi$ unsuccessful, and (obviously) vice versa. For example, even though the negation of that formula is obviously unsuccessful, $\neg(p \wedge \neg K_n p)$ is successful. Instead of a direct proof, it suffices to observe that $\neg(p \wedge \neg K_n p)$ is equivalent to $\neg p \vee K_n p$, and that that formula is in the language fragment that is preserved under taking arbitrary submodels, and therefore *a fortiori* under the unique submodel resulting from its announcement. This is generalized in Proposition 8 (see below).

Finally φ and ψ may be successful but $[\varphi]\psi$ not. Consider a model M with $\{s, t\}$ as the set of possible worlds. There is only one accessibility relation $\sim_a = \{(s, s), (s, t,), (t, s), (t, t)\}$ and only one propositional variable p, which is only true in t, i.e. $V_p = \{t\}$. We take the epistemic state (M, s). Now consider the formula $[\neg p \rightarrow K_a \neg p]\bot$. The subformulas $\neg p \rightarrow K_a \neg p$ and $\bot$ are both successful. However $(M, s) \models \langle[\neg p \rightarrow K_a \neg p]\bot\rangle\neg[\neg p \rightarrow K_a \neg p]\bot$. This can be seen as follows. The formula $\neg p \rightarrow K_a \neg p$ is true in t, but false in s. Therefore $[p \rightarrow K_a p]\bot$ is trivially true in s. It is obviously false in t. So M restricted to this formula consists of s only. In this model $\neg p \rightarrow K_a \neg p$ is true. Therefore $\langle\neg p \rightarrow K_a \neg p\rangle\top$, which is equivalent to $[\neg p \rightarrow K_a \neg p]\bot$, is true there as well.

There are some results. First, *common knowledge formulas* are successful:

PROPOSITION 7. (van Ditmarsch (2003)) . Let $\varphi \in \mathcal{L}_N(P)$. Then $[C\varphi]C\varphi$ is valid.

Proof. Let M, s be arbitrary. Observe that $M, s \models C\varphi$ implies $M|C\varphi, s \models C\varphi$: the truth of a proposition is determined by the set of N-accessible states.[3] But that implication is the simple result of applying the semantical definition to $M, s \models [C\varphi]C\varphi$. □

By announcing a common knowledge formula, no accessible states are deleted from the model. Obviously the truth of formulas can only change by an announcement if their truth value depends on states that are deleted by the announcement. We will now show that formulas from the following large fragment $\mathcal{L}_N^{u0}$ (of the logical language $\mathcal{L}_N^u$ assumed throughout) of the *preserved formulas* with inductive definition

$$\varphi ::= p \mid \neg p \mid \varphi \wedge \psi \mid \varphi \vee \psi \mid K_n\varphi \mid C\varphi \mid [\neg\varphi]\psi$$

are truth preserving under 'deleting states'. From this, it also follows that the fragment is successful. Instead of 'deleting states', we say

that we restrict ourselves to a *submodel*: a restriction of a model to a subset of the domain, with the obvious restriction of access and valuation to that subset.

PROPOSITION 8. (*preservation*). Fragment $\mathcal{L}_N^{u0}$ is preserved under submodels.

Proof. By induction on $\mathcal{L}_N^{u0}$. The case for propositional variables, conjunction, and disjunction is straightforward.

Let $M = \langle S, \sim, V\rangle$ be given and let $M' = \langle S', \sim', V'\rangle$ be a submodel of it. Suppose $s \in S'$. Suppose $(M, s) \models K_n\varphi$. Let $s' \in S'$ and $s \sim_n s'$. Therefore $s \sim_n s'$, and so by the semantics we have $(M, s') \models \varphi$. Therefore, by the induction hypothesis $(M', s') \models \varphi$. Therefore $(M', s) \models K_n\varphi$. The case for $C\varphi$ is completely analogous.

Suppose $(M, s) \models [\neg\varphi]\psi$. Suppose, towards a contradiction, that $(M', s) \not\models [\neg\varphi]\psi$. Therefore, by the semantics, $(M', s) \models \neg\varphi$ and $(M'|\neg\varphi, s) \not\models \psi$. Therefore, by using the contrapositive of the induction hypothesis, also $(M, s) \models \neg\varphi$. Moreover $M'|\neg\varphi$ is a submodel of $M|\neg\varphi$, because a state $t \in S'$ only survives if $(M', t) \models \neg\varphi$, therefore by the induction hypothesis $(M, t) \models \neg\varphi$. So $S'|\neg\varphi \subseteq S|\neg\varphi$. But from $(M, s) \models [\neg\varphi]\psi$ (which we assumed) and $(M, s) \models \neg\varphi$ follows $(M|\neg\varphi, s) \models \psi$, therefore by the induction hypothesis also $(M'|\neg\varphi, s) \models \psi$. This contradicts our earlier assumption. Therefore $(M', s) \models [\neg\varphi]\psi$. □

For a similar fragment, namely without common knowledge and without any announcement operators, this was proved in van Benthem (2002). However, also the converse held for that fragment, i.e. if a formula of the language is preserved under submodels, then it is in the fragment. It is unknown whether this also holds for $\mathcal{L}_N^{u0}$.

COROLLARY 9. Let $\varphi \in \mathcal{L}_N^{u0}$ and $\psi \in \mathcal{L}_N^{u}$. Then $\varphi \to [\psi]\varphi$ is valid.

COROLLARY 10. Let $\varphi \in \mathcal{L}_N^{u0}$. Then $\varphi \to [\varphi]\varphi$ is valid.

COROLLARY 11. (*all preserved formulas are successful*). Let $\varphi \in \mathcal{L}_N^{u0}$. Then $[\varphi]\varphi$ is valid.

We have found successful formulas outside $\mathcal{L}_N^{u0}$, such as $\neg K_n p$, see above. There are more successful formulas than preserved formulas, because the entailed requirement that $\varphi \to [\psi]\varphi$ is valid *for arbitrary*

ψ is much stronger than the requirement that $\varphi \to [\varphi]\varphi$ is valid. In the last case we are only looking at the very specific submodel resulting from the announcement of *that* formula, not at arbitrary submodels.

The result formulated in Proposition 8 is rather like a result of Gerbrandy (1999, pp. 100–101). He proved, for a slightly different notion of (successful) updates, and a language without common knowledge, that formulas are successful if epistemic formulas do not occur within the scope of an odd number of negations.

A syntactic characterization of the whole successful fragment of the language is still not found. We hope to continue making progress on that.

9. CONCLUSIONS AND FURTHER RESEARCH

We have focused on formulas that, when announced, become false: unsuccessful updates. The appropriate logic to investigate this phenomenon is the logic of public announcements, and part of our contribution is to give clear concepts concerning successful formulas and successful updates and some of their elementary semantic properties. We presented four case studies involving unsuccessful updates, that have not previously been presented together with this particular focus in mind.

We hope that this paper will induce others to alert us to other occurrences of unsuccessful updates in particular in the philosophical literature. We surmise that, even apart from 'Moore-problems', many others may have struggled with these matters in the past. Actually, even in the area of 'mathematical recreation' there are examples that we have chosen to overlook, because of their mainly combinatorial interest. For example, the famous 'sum and product'-riddle, where two persons communicate to each other, their ignorance and knowledge concerning the sum and product of two natural numbers that they have been told, masterfully uses the power of unsuccessful updates. For that, see Freudenthal (1969), or for an epistemic treatment, Plaza (1989).

One important technical concern that we hope to resolve in the near future is the syntactic characterization of successful formulas. Beyond that, there are some generalizations of the concepts of successful and unsuccessful execution: an announcement is just one of many conceivable epistemic actions. A formula that when announced becomes false, is in other words a formula that cannot be

announced twice in succession. A generalization of that is the action that cannot be executed twice. An action cannot be executed twice if its precondition is an unsuccessful update.

We hope that the underlying investigation and our continued efforts to enhance our understanding will show the enduring relevance of this subject matter for philosophical and epistemological investigations.

NOTES

[1] The proof system in Table I is complete if the logical language is extended with action composition and the standard PDL axiom for composition. This issue is beyond the scope of this paper.

[2] Erik Krabbe pointed out to us that Gamow and Stern made a slight error and should have said action suddenly broke out on the 39th day instead of the fortieth.

[3] It seems difficult to be more precise here without being cumbersome. Let's make an attempt after all: Given the assumption $M, s \models C\varphi$, observe that $s \sim_N s'$ implies $M, s' \models C\varphi$. In other words: $[s]_{\sim_N} = \{s' \in \mathcal{D}(M) \mid s \sim_N s'\} \subseteq \{s' \in \mathcal{D}(M) \mid M, s' \models C\varphi\}$. But that's another way of saying that $M|[s]_{\sim_N} \subseteq M|C\varphi$ – where we use $M|[s]_{\sim_N}$ par abus de langage in its obvious meaning of '*N-reduced* model'. Therefore, given the assumption, $M, s \models C\varphi$ is equivalent to $M|[s]_{\sim_N}, s \models C\varphi$ is equivalent to $M|C\varphi, s \models C\varphi$.

REFERENCES

Alchourrón C. E., P. Gärdenfors and D. Makinson: 1985, 'On the Logic of Theory Change: Partial Meet Functions for Contraction and Revision', *Journal of Symbolic Logic* **50**, 510–530.

Baltag, A., L. Moss and S. Solecki: 2003, 'The Logic of Public Announcements, Common Knowledge and Private Suspicions', Manuscript, originally presented at TARK 98.

Barwise, J.: 1981, 'Scenes and other Situations', *Journal of Philosophy* **78**(7), 369–397.

Fagin, R., J. Halpern, Y. Moses and M. Vardi: 1995, *Reasoning About Knowledge*, MIT Press, Cambridge, Massachusetts.

Freudenthal, H.: 1969, 'Stating of the Sum-and-Product Problem', *Nieuw Archief voor Wiskunde (New Archive of Mathematics)* **17**, 152.

Gamow, G. and M. Stern: 1958, *Puzzle-Math*, Macmillan, London.

Gärdenfors P.: 1988, *Knowledge in Flux, Modeling the Dynamics of Epistemic States*, MIT Press.

Gerbrandy, J.: 1999, *Bisimulations on Planet Kripke*, Ph.D. thesis, University of Amsterdam. ILLC Dissertation Series DS-1999-01.

Hintikka, J.: 1962, *Knowledge and Belief, An Introduction to the Logic of the Two Notions*, Cornell University Press, Ithaca & London.

Kahn, C. H.: 1979, *The Art and Thought of Heraclitus*, Cambridge University Press, Cambridge.
Kirkman, T.: 1847, 'On a Problem in Combinations', *Cambridge and Dublin Mathematics Journal* **2**, 191–204.
Kvanvig, J. L.: 1998, 'Paradoxes, Epistemic', In E. Craig (ed.), *Routledge Encyclopedia of Philosophy*, Vol. 7, Routledge, London, pp. 211–214.
Makarychev, K. and Y. Makarychev: 2001, 'The Importance of Being Formal', *Mathematical Intelligencer* **23**(1), 41–42.
McCarthy, J.: 1990, 'Formalization of Two Puzzles Involving Knowledge', In V. Lifschitz (ed.), *Formalizing Common Sense: Papers by John McCarthy, Ablex Series in Artificial Intelligence*, Ablex Publishing Corporation, Norwood, NJ. Available online at http://www-formal.stanford.edu/jmc/.
Moses, Y. O., D. Dolev and J. Y. Halpern: 1986, 'Cheating Husbands and other Stories: A Case Study in Knowledge Action and Communication', *Distributed computing* **1**(3), 167–176.
O'Connor, D. J.: 1948, 'Pragmatic Paradoxes', *Mind* **57**, 358–359.
Plaza, J.: 1989, 'Logics of Public Communications', In M. Emrich, M. Pfeifer, M. Hadzikadic and Z. Ras (eds.), *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, 201–216.
Quine, W. V. O.: 1953, 'On a So-called Paradox', *Mind* **62**, 65–67.
Scriven, M.: 1951, 'Paradoxical Announcements', *Mind* **60**, 403–407.
Sorensen, R. A.: 1988, *Blindspots*, Clarendon Press, Oxford.
van Benthem, J. F. A. K.: 2002, 'One is a Lonely Number': On the Logic of Communication', To be published in the *Proceedings of the Logic Colloquium 2002*.
van Ditmarsch, H. P.: 2000, *Knowledge games*, Ph.D. thesis, University of Groningen. ILLC Dissertation Series DS-2000-06.
van Ditmarsch, H. P.: 2002, 'The Description of Game Actions in Cluedo', In L. Petrosian and V. Mazalov (eds.), *Game Theory and Applications*, Vol. 8, Nova Science Publishers, Commack, NY USA, pp. 1–28.
van Ditmarsch, H. P.: 2003, 'The Russian Cards Problem'. *Studia Logica* **75**, 31–62.
van Ditmarsch, H. P., W. van der Hoek and B. P. Kooi: 2005, 'Dynamic Epistemic Logic', Manuscript.
Weiss, P.: 1952, 'The Prediction Paradox', *Mind* **61**, 265–269.

Hans van Ditmarsch
Computer Science, University of Otago,
PO Box 56, Dunedin 9015,
New Zealand
E-mail: hans@cs.otago.ac.nz

Barteld Kooi
Philosophy, University of Groningen,
A-weg 30, 9718 CW Groningen,
The Netherlands
E-mail: barteld@philos.rug.nl